**سوالات تمرين شماره يک BIG DATA**

تاريخ تحويل : 2/9/1401 نمره : 1.5

1. فايل JaneAusten.txt شامل کليه نوشته هاي جین آستین، نویسنده انگلیس، مي باشد که توسط پروژه گوتنبرگ جمع آوری شده است. با استفاده از SPARK برنامه اي بنويسيد که،

* الف: تعداد کل کلمات اين متن را مشخص نماييد.
* ب: تعداد کلمات بدون تکرار چقدر است؟
* ج: ده کلمه اي که بيشترين تعداد تکرار را دارند کدامند و هر کدام چند بار تکرار شده اند؟ همچنین با تغيير تعداد هسته هايي که برنامه (ب) بر روي انها اجرا مي شود از يک هسته تا حداقل چهار هسته نمودار زمان اجراي ان را رسم نماييد. خروجي گزارش شامل کد، نتايج، نمودار و تحليل ان مي باشد.

1. سه فايل C1,C2,C3 هر کدام حاوي اطلاعات دو ستوني متني مي باشند که هر سطر ان مختصات يک نقطه دو بعدي را نشان مي دهد. با استفاده از SPARK دو برنامه بنويسيد که با استفاده از دو روش خوشه بندي K-means++ و Bisecting K-means،

* الف : براي هر فايل داده نمودار هزينه کلاستر را براي k=2 to25 رسم نماييد.
* ب: تعداد بهينه خوشه ها چقدر است؟
* ج: در تعداد بهينه خوشه و در روش خوشه بندي با هزينه کمتر نقاط مرکزي هر کلاستر را براي هر يک از فايل هاي داده مشخص نماييد؟
* د: درتعداد بهينه خوشه نمودار زمان اجراي دو روش خوشه بندي را بر روي يک تا 4 هسته براي هر يک از فايل ها رسم نموده و تحليل نماييد. گزارش شامل کد و نتايج و نمودارهاي هر مرحله و تحليل جواب ها مي باشد.